Sparse- vs. Dense Retrieval for Climate claim Verification

Group 1
Dominik Kolak, Jonas Krauzer

***Abstract*—In this report we compare sparse (BM25) and dense (transformer) retrieval methods for climate claim verification on the Climate-FEVER dataset. The Sparse Approach uses BM25 with RoBERTa-MNLI classification, while our dense Approach uses sentence transformers with Climate-BERT re ranking and classification. Overall we can report that the Dense Approach achieved a 34 % improvement in Accuracy compared to the sparse Approach. We also identified that the used generic MNLI model heavily over predicts REFUTES by a factor of 3.9, while the domain specific model achieved more balanced predictions. Dense retrieval showed a marginal advantage in recall at low k, but this gap widened at higher k, up to a advantage of 6.8 % at k= 100.**

I. Introduction

Climate change is a major threat to our society. While the 2025 United Nations Climate Change Conference did not yield satisfying results, and the 1.5-degree becomes increasingly uncertain, and fake news spreads over the internet. Because of that, we will build a Citation-Grounded Retrieval system for Fact-Checking, which allows the user to enter a claim and search for evidences that prove or falsify the given statement. This evidences are re ranked and classified, returning one of the values Supports, Refutes, or NotEnoughInformation for the entered statement.
We will compare two different approaches:

- Approach 1: uses sparse retrieval with BM25 and a BERT-based classifier
- Approach 2: uses dense retrieval with sentence transformers and a domain specific Climate-BERT classifier

For Approach 1, we expected a better runtime and less implementation effort. Approach 2 is expected to yield better results.

Research question: How much can we improve climate-claim verification by replacing sparse retrieval with dense retrieval and using a domain specific classifier?

II. Related work and Project discussion

A. FEVER and Climate-FEVER

The Fact Extraction and Verification is a framework for automated fact-checking using evidences from Wikipedia. FEVER involves retrieving relevant documents and extracting evidence sentences that are then classified. Climate-FEVER is a adaptation of this framework for climate science claims. This adaptation of FEVER is especially challenging due to the prevalence of misinformation and the specialized terminology. Claims in the Dataset are short and easy to read, resembling real world statements found on the internet, media and news.

B. Climate-FEVER Dataset

Our work is based on the Climate-FEVER dataset, which was introduced in [1]. The dataset consists out of 1535 real-world claims about climate change, each claim containing five evidences, which may support or refute the evidence. The claims were collected from scientifically-informed and skeptics/climate deniers sources, leading to a balanced set of claims. The paper also introduced the term of verifiable claims: A claim is potentially verifiable, if it is

- consistent, unambiguous and not too much implicit knowledge is required)
- the document collection contains knowledge, that helps to support or refute the claim

The evidences were extracted from Wikipedia using a three-step-pipeline (Document level retrieval, Sentence-level retrieval and Sentence re ranking). Every claim-evidences pair was manually annotated by experts.

C. Project discussion

For our project, we process the existing Climate-FEVER dataset further: We extract claims, evidences and the mapping from the initial dataset and reconstructed the claim verification using two approaches. This allowed us to compare the effectiveness of sparse and dense retrieval methods combined with generic and domain specific classifiers. Our approach differs from the original FEVER paper in several ways. We focused on the Retrieval Method instead of the Overall pipeline like in FEVER. The pretrained models we used are also more recent then the ones used in FEVER since these models weren't available at the time. We designed two very contrasting approaches, one using a generic model with sparse retrieval and the other a domain specific model with dense retrieval to maximize the insight into what components provide the highest contribution to this pipeline.

D. Evaluation Methodology

We evaluated both approaches on a stratified 80/20 validation split with a fixed random state. This stratification prevented label biases by maintaining a equal label distribution. For the retrieval evaluation, we used:

- Recall@k: Proportion of ground-truth evidences found in top-k results.

```
{
  "claim_id": "5",
  "claim": "The sun has gone into lockdown
      which could cause freezing weather,
      earthquakes and famine, say scientists
      ",
  "claim_label": "SUPPORTS",
  "evidences": [
    {
      "evidence_id": "Famine:386",
      "evidence_label": "SUPPORTS",
      "article": "Famine",
      "evidence": "The current consensus of
          the scientific community is that
          the aerosols and dust released
          into the upper atmosphere causes
          cooler temperatures by preventing
          the sun's energy from reaching the
           ground.",
      "entropy": 0,
      "votes": [
        "SUPPORTS",
        "SUPPORTS",
        null,
        null,
        null
      ]
    },
  ]
}
```


Fig. 1. Example of a FEVER-style claim. In the dataset, five evidences per claim are provided.

- Mean Reciprocal Rank: The average inverse rank of the first relevant evidence

For the classification evaluation, we used:

- Accuracy: Correct Predictions
- Macro F1: Average of per class F1 Score
- Weighted F1: Class Weighted average per class F1 Score
- Per-Class: Precision, Recall, F1
- Confusion Matrices to identify Biases

These metrics were computed with scikit-learn evaluation functions.

III. Experiments and Results

A. Data preparation

We applied text normalization to all claims and evidence passages. These normalizations included Unicode normalization with NFKC and mapping of special characters to ASCII equivalents were done with regular expressions. The original dataset was restructured into three relational tables: claims.csv containing unique claims with their labels, evidences.csv containing de duplicated evidence passages with articles, and mappings.csv with the claim-evidence relationships. We filtered out low quality evidence passages that were too short (< 40 characters), mostly numeric ($> 50\%$ digits), or contained only citations. After preprocessing, we retained 1535 claims and 5485 unique evidence passages for evaluation.

B. Approach 1: Sparse Retrieval with Generic Classifier

1) Retrieval: We used a BM25 retriever to get the top-50 results for each claim and english stemming to improve accuracy. BM25 scored the evidences with TF-IDF based on how often claim terms appeared in the evidences. This penalized terms that appeared frequently in the corpus but also rewarding relevance. Stemming reduced terms to their root, this improved accuracy by matching related words better. Stemming was used for both the corpus and the claims during retrieval.

2) Classification: Claim verification was performed with roberta-large-mnli, a RoBERTa-large model fine-tuned on the Multi-Genre Natural Language Inference Dataset. We evaluated each of the 50 retrieved evidences by pairing them with the target claim. The model treated this as a natural Language Inference task and mapped the entailment, contradiction and neutral outputs to SUPPORTS, REFUTES and NOT ENOUGH INFO labels.

3) Aggregation: To produce the final result we implemented majority voting and weighted voting. We ultimately ended up using a weighted approach, summing the models confidence scores for each label for the evidences. To improve reliability we introduced a override of the label in the case that any confidence is present for either SUPPORT or REFUTE, in this case NOT ENOUGH INFO was excluded. Similarly we introduced a threshold that would require the leading label to exceed 60 % confidence or it would otherwise default to NOT ENOUGH INFO.

C. Approach 2: Dense Retrieval with Domain Specific Classifier

1) Retrieval: We included dense information retrieval using the all-MiniLM-L6-v2 sentence transformers to create 384-dimensional embeddings for candidate evidence selection. Unlike BM25 this approach captured semantic meaning and identified relevant evidences even when lexical similarity was low. We precomputed normalized embeddings for the evidence corpus and for testing purposes we did the same for the all the claims. Since we were working with normalized vectors we could efficiently compute similarity with the dot product. For each claim we retrieved the top 100 candidates to have broad coverage and increase the recall as much as possible.

2) Rerank: Since the initial retrieval was so broad we used a Climate-BERT cross-encoder as a re ranker. The initial bi-encoder embeds the claims and evidences separately to increase retrieval speed. The cross-encoder on the other hand processes claim-evidence pairs together. We computed the relevance scores as P(SUPPORTS) + P(REFUTES) while excluding NOT ENOUGH INFO. This was done to prioritize decisive evidences. Afterwards we selected the top-10 best scoring evidences for the final classification.

3) Classification and Aggregation: We reused our Climate-BERT model to classify each of the top-10 re ranked evidences with our labels. With this approach we could ensure that both the classification and the re ranking had the same semantic understanding of the evidences. The final result was reached

with weighted voting and the same restrictions of priority and confidence as in approach 1.

D. Results - Retrieval

As supposed, dense retrieval yielded a better recall score. But as we can see in table I the difference for k $\leq$ 10 was tiny and increased only slightly for a higher k. This benefit came with a higher performance cost for the creation of the embeddings. Consistently, the Mean Reciprocal Rank (MRR)

TABLE I
RECALL@K FOR THE DIFFERENT APPROACHES IN PERCENT

k	BM25	Dense	Delta
5	32.8	32.9	0.1
10	42.4	42.6	0.1
20	51.1	54.1	3.1
50	62.8	68.9	6.1
100	72.0	78.7	6.8

was slightly higher for the dense retrieval, as table II shows. Setting k = 50, Approach 1 found all evidences for 405

TABLE II
MRR FOR BOTH APPROACHES

	BM25	Dense	Delta
MRR	0.38	0.40	0.02

claims, but did not find any matching evidence for 190 claims. Approach 2 could find all evidences for 50 more claims. The complete histograms can be seen in figure 3 and 4. Looking at claim 31, we can see an example where dense retrieval performed superior to BM25:

"Discovery Of Massive Volcanic CO2 Emissions Discredits Global Warming Theory"

For this claim, BM25 could not find a single evidence because only the very general words *warming*, *global* and *of* are in both texts (even stemming does not help here), whereas dense retrieval found all three:

- *"Intrusions of hot magma into carbon-rich sediments may have triggered the degassing of isotopically light methane in sufficient volumes to cause global warming and the observed isotope anomaly."*
- *"The eruptions would also have emitted carbon dioxide, causing global warming."*
- *"The basalt lava erupted or intruded into carbonate rocks and into sediments that were in the process of forming large coal beds, both of which would have emitted large amounts of carbon dioxide, leading to stronger global warming after the dust and aerosols settled."*

E. Results - Classification

For the classification, we calculated the macro- and weighted f1-score for each approach. We can see that Approach 2 yields better scores. This is the supposed result, as Approach 2 scored better in the retrieving part and uses a specialized classifier (ClimateBERT).

TABLE III
ACCURACY AND F1 SCORES

Metric	Approach 1	Approach 2	Delta
Accuracy	0.32	0.65	0.33
F1 (macro)	0.31	0.59	0.28
F1 (weighted)	0.29	0.63	0.34

TABLE IV
F1-SCORE PER CLASS

Class	Approach 1	Approach 2
SUPPORTS	0.21	0.68
REFUTES	0.34	0.41
NOT_ENOUGH_INFO	0.33	0.68

The classification scores for Approach 1 can be obtained from table VII. We can see that approach 1 classifies most claims with the SUPPORTS label wrong. This behavior could be reproduced in several test runs (See also figure 6).

F. Classification Bias Analysis

Approach 1 shows a significant REFUTES bias. Of 131 SUPPORTS claims only 17 roughly 17% were correct while 61% were misclassified as REFUTES. This approach over predicted REFUTES 157 times against the ground truth, over predicting by a factor of 3.1.

TABLE V
CONFUSION MATRIX FOR APPROACH 1

	SUPPORTS	REFUTES	NOT_ENOUGH_INFO
SUPPORTS	17	61	53
REFUTES	2	38	11
NOT_ENOUGH_INFO	2	58	34

TABLE VI
CONFUSION MATRIX FOR APPROACH 2

	SUPPORTS	REFUTES	NOT_ENOUGH_INFO
SUPPORTS	102	10	19
REFUTES	24	16	11
NOT_ENOUGH_INFO	35	10	49

This bias comes from the generic MNLI lacking domain specific understanding and BM25 retrieving results of mixed quality. The weighted voting only increased these contradictory results. Approach 2 shows a far more balanced result but even here REFUTES remains suboptimal (31% Recall). This can be either due to the class imbalance for REFUTES which was 18.5% or can be explained by the semantic difficulty in distinguishing a REFUTES to a NOT ENOUGH INFO.
The detailed scores show the tradeoff, Approach 1 achieved high precision for SUPPORTS (81%) but terrible recall (13%), while REFUTES shows the opposite (24% Precision, 75% Recall). This shows that the model is accurately predicting SUPPORTS but rarely makes this prediction, instead defaulting to REFUTES for most claims. Again Approach 2 produces more balanced predictions with SUPPORTS achieving

reasonable results for both precision (63%) and good recall (78%). While REFUTES drops to 31%. REFUTES remains a issue even in this more sophisticated approach due to the class imbalance and the inherent semantic difficulty in distinguishing REFUTES from NOT ENOUGH INFO.

TABLE VII
CLASSIFICATION SCORES FOR APPROACH 1

	precision	recall	f1-score	support
NOT_ENOUGH_INFO	0.35	0.36	0.35	94
REFUTES	0.24	0.75	0.37	51
SUPPORTS	0.81	0.13	0.22	131

TABLE VIII
CLASSIFICATION SCORES FOR APPROACH 2

	precision	recall	f1-score	support
NOT_ENOUGH_INFO	0.62	0.52	0.57	94
REFUTES	0.44	0.31	0.37	51
SUPPORTS	0.63	0.78	0.70	131

G. Runtime Comparison

We measured the total execution time for processing 1378 Claims. Approach 1 completed the entire set in 13:18 Minutes, averaging 0.57 seconds claim. Compared to Approach 2 with a total runtime of 16:43 Minutes and a average runtime per claim of 0.72 seconds. We can observe that Approach 2 takes approximately 26% longer to process a equal subset.

Even though Approach 1 is significantly simpler we weren't able to observe the expected runtime benefit for the approach compared to the more sophisticated dense passage retrieval approach with re ranking.

TABLE IX
F1-SCORE PER CLASS

	Runtime total	Runtime per Claim
Approach 1	13m 18s	0.57s
Approach 2	16m 43s	0.72s

IV. CONCLUSIONS AND FUTURE WORKS

In our project, we could show that a simple approach with BM25 and a simple BERT-based classifier has slightly lower computation with the drawback of significantly lower scores compared to an approach with dense retrieval (DR) and a domain specific BERT-classifier. Though our DR-pipeline shows improved metrics, but the results are still not satisfactory and might be able to be further improved using a hybrid approach for retrieval with a better dataset. A better classifier model for re ranking and voting could further improve the results.

REFERENCES

[1] Thomas Diggelmann et al. "CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims". In: *CoRR* abs/2012.00614 (2020). arXiv: 2012.00614. URL: https://arxiv.org/abs/2012.00614.

V. APPENDIX

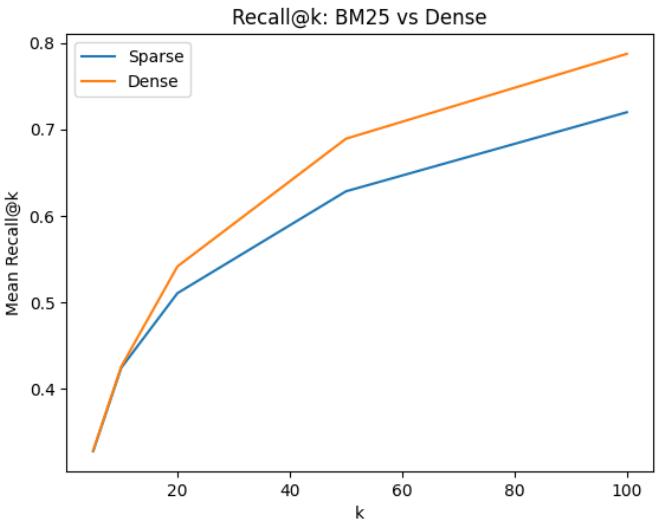

Fig. 2. Recall for Approach 1 (Sparse retrieval) and Approach 2 (Dense retrieval)

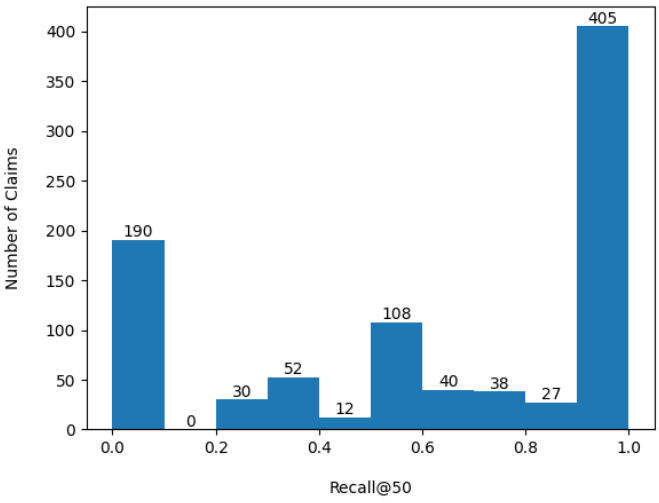

Fig. 3. Histogram of Recall scores for Approach 1

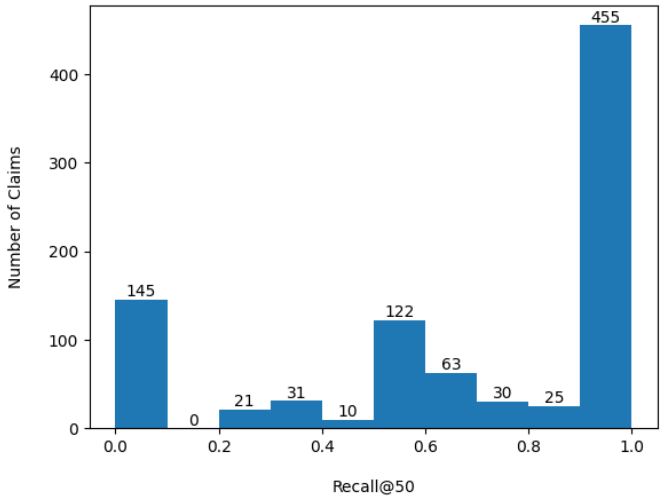

Fig. 4. Histogram of Recall scores for Approach 2

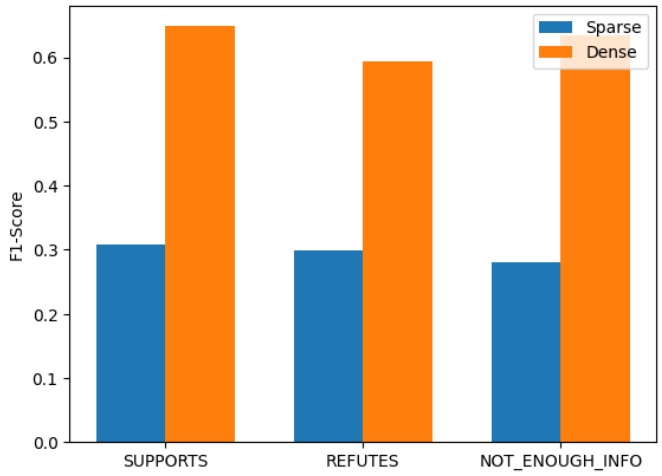

Fig. 5. Comparison of F1-score

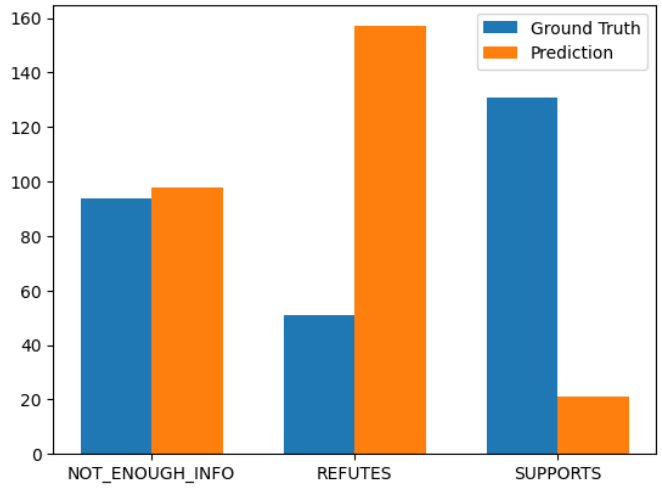

Fig. 6. Comparison of the amount of predicted labels (Approach 1)

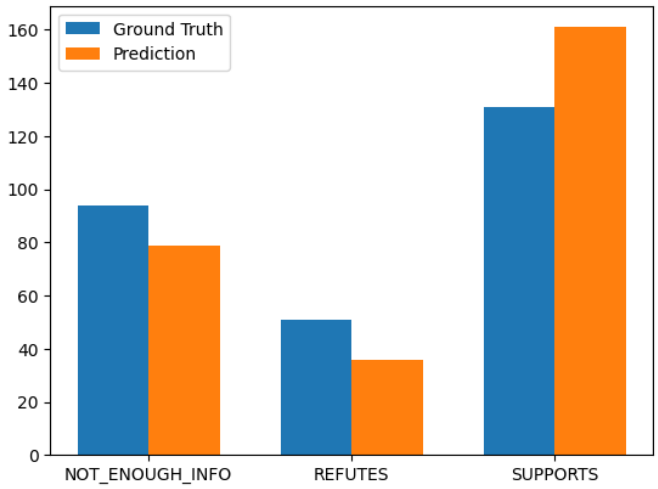

Fig. 7. Comparison of the amount of predicted labels (Approach 2)